A METHOD FOR DEVELOPING INFORMATION RETRIEVAL SYSTEMS FOR SMALL LIBRARIES

by

Eddie Lorenz Zeitler

A Thesis Submitted to the Faculty of the

DEPARTMENT OF SYSTEMS ENGINEERING

In Partial Fulfillment of the Requirements
For the Degree of

MASTER OF SCIENCE

In the Graduate College

THE UNIVERSITY OF ARIZONA

1 9 7 0

STATEMENT BY AUTHOR

This thesis has been submitted in partial fulfillment of requirements for an advanced degree at The University of Arizona and is deposited in the University Library to be made available to borrowers under rules of the Library.

Brief quotations from this thesis are allowable without special permission, provided that accurate acknowledgement of source is made. Requests for permission for extended quotation from or reproduction of this manuscript in whole or in part may be granted by the head of the major department or the Dean of the Graduate College when in his judgment the proposed use of the material is in the interests of scholarship. In all other instances, however, permission must be obtained from the author.

SIGNED: Eddie J. Tatt

APPROVAL BY THESIS DIRECTOR

This thesis has been approved on the date shown below:

Robert L. Baker
ROBERT L. BAKER
Assistant Professor of Systems Engineering

May 16, 1970
DATE

PREFACE

In June, 1969, I was contacted by Dr. Robert Baker, Assistant Professor of Systems Engineering at The University of Arizona, concerning the development and implementation of an information retrieval system for the Office of Arid Lands Studies. This proposed system afforded an excellent opportunity to apply systems engineering techniques in the field of Information Science. There have been relatively few applications of this nature and I was pleased to have been afforded the opportunity to work on this project. This paper contains the work performed in developing the CANED System.

The applications of systems engineering presented herein are by no means complete. Further areas warranting continued research are (1) the evaluation of the performance of the CANED System and (2) the development of an expanded system, incorporating the six arid lands research centers of the western United States into a cooperative information network.

I should like to thank Capt. Thomas Canavan, USAF, for his invaluable help with the writing of the computer programs, Dr. Robert Baker and Patricia Paylore for their advice and guidance, the two members of my committee, Dr. Wymore and Professor Perry, and my wife, Vickie, for her unfailing moral support.

TABLE OF CONTENTS

LIST OF ILLUSTRATIONS

ABSTRACT

This paper discusses the design procedure employed in the development of the CANED Information Retrieval System for the Office of Arid Lands Studies of The University of Arizona. This procedure involved the application of systems engineering techniques to design methodology, utilizing a model proposed by the Comparative Systems Laboratory at Western Reserve University which incorporates the basic components of an information retrieval system as a guideline. Chapter I describes the steps comprising the design process of an information system. Chapter II provides a survey of current literature in the field of information science. In Chapter III, the actual development of the CANED System is treated, and Chapter IV discusses the possible future developments of the system.

CHAPTER I

THE ESTABLISHMENT OF A DESIGN PROCEDURE

Introduction

Ralph Shaw (109) has questioned the value of most automated information retrieval systems. He stated that " . . . the claim that machine search is 'above all cheaper' could make sense only if we were paying librarians and other bibliographers about $2000 an hour." Continuing, he implied that no one is able " . . . to name a single application of the computer to information retrieval for which it can be shown that the computer is currently accomplishing anything of significance that cannot be done faster and more cheaply by hand." He is perhaps a bit over zealous, but he does raise some very valid questions concerning the appropriateness, in many cases, of automated retrieval systems. Historically, automated retrieval systems have been designed for very large data bases and are, necessarily, extremely sophisticated and costly. The cost of many of these large systems can be met due to their extensive use. Unfortunately, there are systems operating where the user's needs are not as sophisticated as the software employed. These systems have a very difficult time justifying their operational costs. For the extremely large library requiring a very costly and sophisticated retrieval system, Mr. Shaw may be right when he says that " . . . we should be putting massive effort into the development of a science upon which a better information technology may be built."

But what of the application of an automated information retrieval system for a small library of say 50,000 to 75,000 documents? This paper discusses the steps in the design of an automated retrieval system to be used with the small technical library. Utilizing the random access capability of a third generation computer, this information system would fulfill the user's needs, yet remain inexpensive enough to justify its existence.

Systems Engineering

The complex nature of information systems creates a challenge for their design and evaluation. A number of authors, including Fasana (36), Goodman, et al. (43), Kilgour (65), Kraft (70), Moore (88), and Stevenson (112), have advocated adopting systems engineering techniques to cope with problems of design and evaluation of information systems. The systems engineering method of problem solving, or systems approach, involves the realization that factoring out part of a problem by neglecting the interactions among its sub-systems and elements increases the probability that a solution to the design problem will not be found. The application of the systems approach requires that the system be planned and designed as an entity in order to satisfy the needs of the user. Each element of the system must be planned and designed by explicitly considering its interrelationships with the other elements of the system, and by evaluating and optimizing its cost and benefit as part of the total system's cost and benefit. Kline and Lifson (69) list three primary considerations that must be included in the design of an effective system:

1) The system attributes that represent fulfillment of needs and objectives--i.e., the design criteria--must be identified and described. In general, there are three primary classes of criteria: (a) effectiveness (benefit) criteria, representing measures of needs fulfillment; (b) resource criteria, representing costs associated with levels of effectiveness; and (c) schedule criteria, representing the time when the system is to be required.

2) The relative importance of each of the design criteria to achievement of program objectives must be evaluated, and weighting relationships among the criteria must be defined.

3) The constraints imposed on the system by its environment must be identified and described.

Saul Herner (49) has stated that modern information systems and their associated concepts have been very much misused. He concludes that the reason " . . . is primarily due to the fact that information retrieval systems have not been designed by people actively involved in the operation and management of library and information programs, but rather, by computer systems specialists, who often try to mold or adjust information program requirements to make them conform to the capabilities of existing hardware of previously designed systems."

In considering the design of an information retrieval system, the systems engineer must realize that both the Computer Systems and Library Science fields are highly specialized. Often a person working in such a specialized area will not fully appreciate the problems encountered in other fields. Therefore, an effective information retrieval system should not be designed by a person who is a specialist in either of these fields. As Dr. A. Wayne Wymore, Head of the Systems Engineering Department at the University of Arizona, points out, a transistor cannot be expected to design the circuit of which it is a part.

The place to start any meaningful discussion of a system, as applied to the collection, organization, and dissemination of information, is by defining the word system. Mr. Herner defines a system as " . . . an ordered, consistent, predictable way of accomplishing a desired end." Of utmost importance is order and consistency in a system. Order and consistency give rise to predictability which is the foundation for dependability and efficient use of a system. Consider, for example, the classification systems used to arrange books in a library. For many years, the Library of Congress classified computers under Calculating Machines, completely ignoring their non-numerical applications. Although this classification may not be exactly correct, it is dependable, or perhaps consistently undependable would be more accurate, and a person could always depend on the fact that books about computers would be shelved with books about calculating machines.

Mr. Herner goes on to state that although a system must be as responsive as possible to the needs of its users, " . . . it can never be completely responsive to the interests, view points, and contexts of all its users; if it were, it would not be a system but an anarchy." Thus, the best system is actually a compromise and the job of the systems engineer is to formulate the best compromise possible. In conclusion then, the design of an effective information retrieval system requires that the designer be aware of the current state of the art of automated information systems and, much more important, be aware of what the system must do to satisfy the needs of the users of that system.

Components of an Information Retrieval System

Before discussing the steps in the design of a system, the identification of the components comprising it should be clearly defined. An information retrieval system has been defined (Saracevic and Rees, 105) as "an integrated assembly of components that interact cooperatively to perform a predetermined function for a specific purpose." In any design or experiment dealing with information retrieval systems, it is possible to specify and distinguish between the purpose components (the what and why of a system) and the function components (the how of a system). The Comparative Systems Laboratory at Western Reserve University (105) has formulated a model incorporating basic components of retrieval systems. This model was used in the design of the CANED Information Retrieval System described in Chapter III. At this time, an explanation of the Comparative Systems Laboratory's model is presented.

The components of the model relating to function and purpose are enumerated thus:

A. Purpose Components
   1) Type of Users
   2) Subject Discipline(s)
   3) File Size

B. Function Components
   1) Acquisition
   2) Indexing Language
   3) Coding
   4) File organization:
      i) Documents
      ii) Document Representations
   5) Question Analysis
   6) Searching Procedures
   7) Dissemination

The elements within a component that contribute to its operation are referred to as factors. The purpose components can be elaborated to illustrate the essential factors that operate within each component.

1) Type of Users
   Identification of User Population
   Specification of User Characteristics
   Establishment of the Size of User Population
   Elaboration of User's Information Needs

2) Subject Discipline(s)
   Definition of the Subject Boundaries (Editorial Scope)
   Establishment of Input Criteria (Inclusions/Exclusions)

3) File Size
   Establishment of Size of File to Yield Satisfactory Output

These factors are interrelated: the formulation of one influences the formulation of others. Most important is the fact that the design of all function components depends on the definition of the purpose components.

The Comparative Systems Laboratory has identified three classes of factors operating within each function component of the system:

1) System Mechanics: Instruments which are incorporated into the system during the operation of the system (i.e., manuals, computer programs, the sauri, etc.).

2) Human Factors: Characteristics of people who apply the system mechanics.

3) System-Human Interaction: Processes by which people apply the system mechanics (i.e., acquisition, indexing, coding or searching).

Although each function component can be described by the three factors given above (system mechanics, human factors, system-human interaction) the Comparative Systems Laboratory took each function component and each factor within a component and further broke them down into elements as shown in Figure 1. It should be noted that there are at least four elements within the System-Human Interaction factor (noted as T.E.M.I.) that operate in all function components:

1) Time Used in Performing Task

2) Environmental Conditions

3) Motivation of Personnel Representing an Amalgam of Personal Motivation and Systems Policies, Rules and Administration

4) Interpretation by Personnel of Systems Roles and Policies (i.e., interpretation of indexing manuals and policies.)

When designing an information system, these basic elements (Figure 1) should be considered. Notice that not all of these elements may be applicable to the particular system under consideration, but the engineer should be aware that they exist and consider each factor as being interdependent within the total structure of the system.

SYSTEM MECHANICS	SYSTEM HUMAN INTERACTION	HUMAN FACTORS
A. Acquisition		
Acquisition Policy	Acquisition Selection	Characteristics of
Sources of Acquisition	Administration	Acquisition Personnel
Acquisition Procedures	T.E.M.I.	(i.e., Subject Background)
B. Indexing Language		
Source of Indexing Language	Indexing:	Indexer Characteristics:
Control of Vocabulary	Editing	Subject Background
Vocabulary Range	Selection of Concepts	Systems Background
Size of Vocabulary	Translation into	Amount of Training
Syntax (roles & links)	Indexing Language	Length of Experience
Relationships Between Terms	T.E.M.I.	Proficiency
Definitions (scope notes)		Personal Characteristic
Degree of Pre-coordination		
Display Format		
Indexing Instructions		
Indexing Policies		

Figure 1: Elements of an Information Retrieval System

SYSTEM MECHANICS	SYSTEM HUMAN INTERACTION	HUMAN FACTORS
Indexing Manuals		
Source of Indexing		
(i.e., Titles, Full Texts, etc.)		
C. Coding		
Representation Selected Notation	Coding:	Coder:
Coding Instructions	Handling of Notation	Characteristics of Coding
Coding Policies	T.E.M.I.	Personnel
Coding Manuals		
D. File Organization		
Mode of Arrangement	Organization of Files	Characteristics of File
File Organization Instructions	Handling of Indexing	Organizing Personnel
File Organization Policies	Information and/or	
File Organization Manuals	Documents	
Recording Media	Handling of Recording Media	
Equipment for Handling of	and Equipment	
Files	T.E.M.I.	

Figure 1: Elements of an Information Retrieval System (Continued)

SYSTEM MECHANICS	SYSTEM HUMAN INTERACTION	HUMAN FACTORS
E. Question Analysis		
Instructions to Questioner	User-Analyst Interaction	Characteristics of Question
Method of Submitting Questions	Enumeration of Search Concepts	Analysis personnel
Analysis Type	Translation into Indexing	
Interaction with Indexing	Language	
Language	T.E.M.I.	
Analysis Instructions		
Analysis Policies		
Analysis Manuals		
Tools for Analysis		
(i.e., Dictionaries)		

Figure 1: Elements of an Information Retrieval System (Continued)

SYSTEM MECHANICS	SYSTEM HUMAN INTERACTION	HUMAN FACTORS
F. Search Procedures		
Available Logic	Selection of Search	Characteristics of Search
Search Parameters	Parameters: (i.e.,	Personnel
Relation of Question	Search Scope	
to Search Parameters	Logic	
Search Method	Generality/Specificity of	
Searching Equipment	Terms	
Search Instructions	Grammer)	
Search Policies	Combination of Parameters into	
Search Manuals	Search Statement	
	Manipulation of Files	
	(searching)	
	T.E.M.I.	

Figure 1: Elements of an Information Retrieval System (Continued)

SYSTEM MECHANICS	SYSTEM HUMAN INTERACTION	HUMAN FACTORS
G. Dissemination		
Dissemination Method	Process of Dissemination	Characteristics of Dissemination
Format of Output	(i.e., Handling of Output,	Personnel
Dissemination Instructions	Reproduction of Documents,	
Dissemination Policies	Delivery of Output to Users)	
Dissemination Manuals	T.E.M.I.	
Administration of Dissemination		
Accompanying Explanation to Users		
Provision of Feedback from Users		

Figure 1: Elements of an Information Retrieval System (Continued)

Design of an Information Retrieval System

By combining the design concepts of systems engineering and the information system components described by the Comparative Systems Laboratory, an information system design procedure was established. This procedure was employed in the construction of the CANED Information Retrieval System discussed in Chapter III. This section discusses the steps invovled in this particular design process.

The first step should be a meeting between the system engineer and the managers of the organization that wish to implement such a system. To establish the need and expectations of the system, three questions should be answered at this time.

1) Why is a system needed?
2) What is it expected to do?
3) What specific problems are expected to be solved?

Often, the motivation behind the desire for a system is not a clear-cut definition of need but a desire to be in the forefront of current fashion. Many times fashions change, leaving a system that resembles a white elephant rather than a viable working mechanism that does the required job in the best and most efficient way.

The next step would be to apply the system considerations given earlier. First, establish the three primary criteria of the system: 1) effectiveness, 2) resources, and 3) schedule. Second, evaluate and define the relative importance of each criteria since it is often the case that they tend to contradict one another. Third, identify the environmental constraints. And finally, review these primary criteria to ensure that they operate within the environmental constraints imposed on the system.

Once a tentative agreement has been made on the design objectives, the designer should consult the probable users and operators of the system. The managers and computer systems people may be highly satisfied with the design and performance of an information retrieval system. However, if the users and operators have not been involved at the design stages, the finished system will often have few applications to the user's needs. This user study is not an attempt to get the users to design the system. Instead, the study will act as a guide in defining the user's real needs and in determining the best available means of meeting them.

Therefore, within the context of the objectives and constraints of the system and the user's immediate needs, the components of an information retrieval system can be evaluated. By matching alternative means against each of these components, a first guess design can be established that has a high probability of acceptance by the users while maintaining a minimal cost.

Having defined and assembled the processes and procedures involved in the system, the next step should be to put them together to make a consistent, dependable system. One device commonly used for arranging the operations of a system is flowcharts. However, whatever method is chosen, the arrangement must be examined to make sure that it is realistic and complete, and that the work flows in a logical, even sequence so that there is a minimum of queuing or delay when proceeding from one process to the next.

The final step in the initial design should be to implement a pilot system and thoroughly test each stage. At this point it is inevitable that something has been done wrong or something has been forgotten. After these problems have been taken care of, and the system is operating satisfactorily, it should be evaluated. Unfortunately, the concept of evaluation in this field is even more misunderstood and misrepresented than the concept of the system itself.

There are many methods available to evaluate a system, and some are more popular than others. However, according to Herner (49), whichever method is chosen, there are five basic questions that must be answered:

1) What services should the system be performing on behalf of its users?

2) Is the system performing all the services it should be performing?

3) Is the system performing any services that it should not be performing?

4) Is it performing the services it should be performing as efficiently and economically as possible?

5) If not, what are the causes, and what can be done to remedy the problem or problems?

Earlier, phrases such as "user's immediate needs," "first guess design," and "system evaluation" were used. These phrases were used because the field of information science is in a very dynamic situation: users' needs change very quickly, and the state of the art in computer science is changing daily. Therefore, there is a need to alter the methods of system evaluation to keep pace with the changing system. Thus, any information retrieval system should be designed so that it can be modified and added to easily. Of course, system evaluation should be a permanent part of the operation of a system if that system is to continue to be a realistic, valuable user tool.

CHAPTER II

A LITERATURE SURVEY

The literature published over the last fifteen years about information science is considerable. With the exploding growth of published material in this field, as with any expanding field, the terminology has become jumbled, often contradictory, and at least confusing to the student beginning his study of information science. There are two texts available, Lancaster (72) and Artandi (4), which serve as adequate introduction to current terminology and concepts. There is also a comprehensive review published yearly by Cuadra (22, 23, 24) in which current developments and trends are presented. This chapter discusses the literature available in four areas of information retrieval:

1) Design of Information Retrieval Systems
2) Subject Indexing, Indexing Terms, and Controlled Vocabularies
3) Search Files and Searching Mechanisms
4) Systems Performance and Evaluation

Design of Information Retrieval Systems

Harold Borko (10) has published an article concerned with the current developments in the design of information retrieval systems. He organized the available literature into the following four catagories:

1) Articles that help define the process of systems analysis and its relationship to information science.

2) Articles that describe the concepts of systems analysis as applied to the analysis and design of information systems.

3) Articles primarily concerned with the design of document storage and retrieval systems.

4) Articles of narrower scope dealing with the design process as applied to subsystems.

In defining the process of systems analysis and its relationship to information science, three government agencies are sponsoring research: the Information Science Directorate of the Air Force Office of Scientific Research (AFOSR), the Information Systems Branch of the Office of Naval Research (ONR), and the National Science Foundation (NSF). Each agency has reported the results of their respective projects in the following publications:

1) Information Sciences, 1965, by Swanson (113) -- a composite of ten studies published by AFOSR.

2) Information Systems Summaries, by Wilcox (125) -- published by ONR.

3) Current Research and Development in Scientific Documentation, No. 14 -- published by NSF (90).

The concepts of systems analysis as applied to the design of information systems is discussed by Kent (63). He devotes a chapter to the problem of specifying a criterion by which the effectiveness of a retrieval system can be calculated. He lists four basic factors to be considered in designing and evaluating systems:

1) Objectives: the overall purpose of the system.

2) Functions: major classes of actions or performance required to achieve the objectives.

3) Performance Requirements: specific dimensions of required actions, with a statement of the standard or required level of performance for each.

4) Environmental Variables: properties of the environment that affect the system and its performance.

Many of the articles on system design are written in a popular rather than a technical style. They are usually aimed at introducing management to systems analysis concepts (i.e., Dearden and McFarlan, 27; McDonough, 84; Prince, 98; Optner, 93; and Quade, 99). For example, in viewing systems analysis as a management tool, Optner (93) lists seven steps to be used in problem solving:

1) The problem process must be flowcharted, showing the principal decision points.

2) Details of the principal decision process steps must be described.

3) The principal alternatives and how they were generated must be fully stated.

4) The assumptions pertinent to each alternative must be identified.

5) The criteria by which each alternative will be judged must be fully stated.

6) Detailed presentation of data, data relationships, and procedural steps by which data were evaluated must be part of any solution.

7) The major alternative solutions, and details to explain why other solutions were eliminated, must be shown.

These steps are compatible with Quade's (99) view of systems analysis as a process " . . . to help a decision maker identify a preferred course of action from among possible alternatives."

Material on systems analysis as an engineering procedure is well covered by Corrigan and Kaufman (21), Wilson and Wilson (127), and Swanson (114). A procedure presented by Corrigan and Kaufman (21) requires that (a) mission objectives and constraints be evaluated, (b) functional flowcharts be prepared, and (c) allocation of specific functions to the man-machine elements operating within the specified constraints and satisfying the mission objectives be made. In reference to (c) above, a discussion of some of the strategies employed in deciding which system functions should be done by man and which by machine are discussed by Chapanis (16) in an article titled "On the Allocation of Functions Between Man and Machines."

The design of document storage and retrieval systems has been discussed by such men as Borko (9), Becker and Hayes (7), and Mayer (82). Figure 2 represents a "scheme for a general model of a document storage and retrieval system" as envisioned by Taulbee (120). This model blocks out the functions of acquisition and data collection, files, searching, and evaluation. It also identifies a set of functions that are expected to be performed by the machine in fully automated systems.

At the 1966 American Documentation Institute convention, one of the disucssion groups addressed the problem of how an information system should be defined. Ronald Wyllys prepared an informal report (128) of the discussion which dealt with some of the basic steps in system design. The major contribution to come from this group was a listing of the major mistakes frequently made in designing systems.

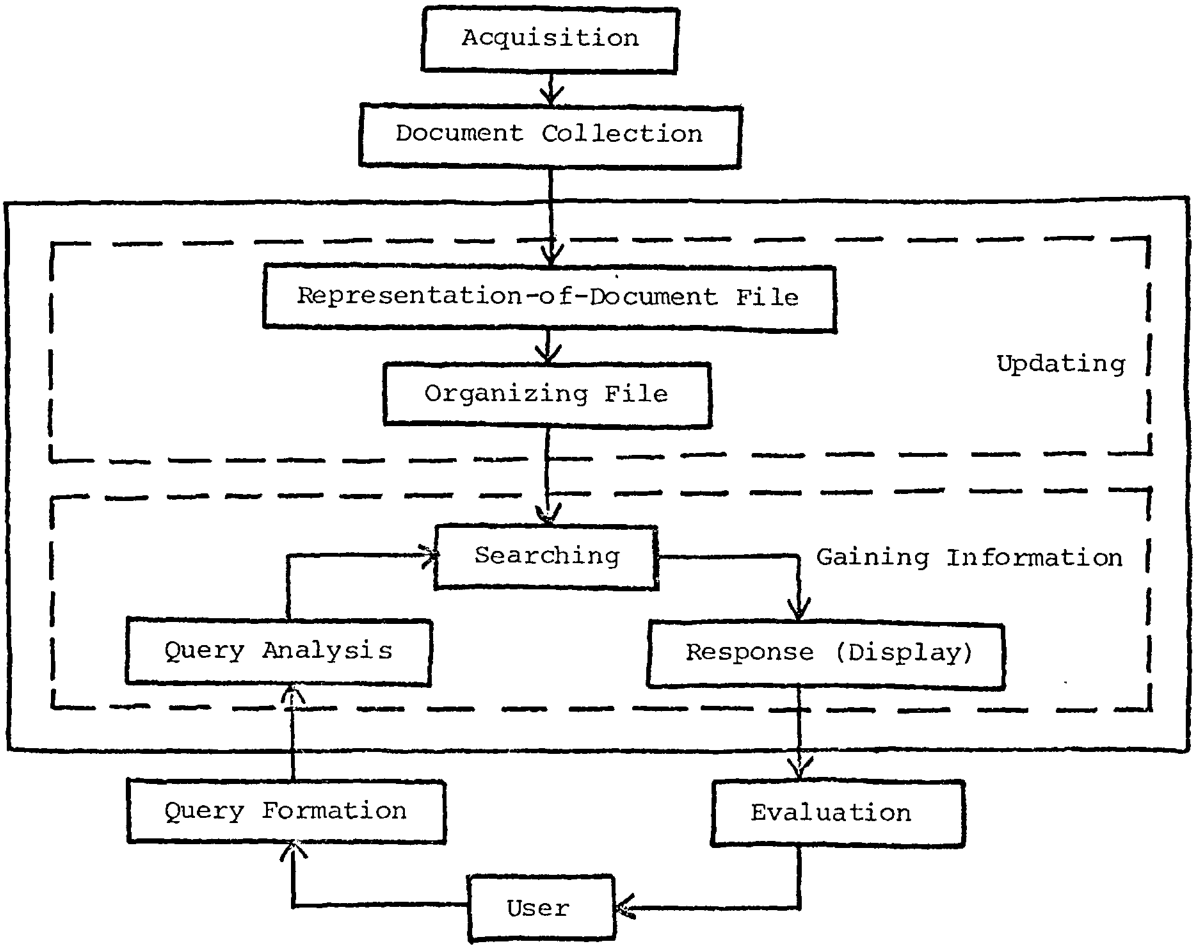

Figure 2. Scheme for a General Model, Taulbee

1) Failing to plan adequately and being overly optimistic with respect to costs, products, and schedules.

2) Failing to recognize changes in the system's goals and characteristics that are necessitated by changes in the system's environment during the course of the design and implementation process.

3) Failing to keep the user involved and informed on the changes that are made in the design of the system.

4) Failing to design the system to accomodate cost analysis data as a prerequisite for evaluation of performance.

The literature reporting on various activities in the design process of system components has been broken into five areas by Borko (10):

1) System objectives and user needs.

2) Methods of document analysis and vocabulary control.

3) File organization and maintenance.

4) On-line retrieval, display, and transmission.

The areas of document analysis and vocabulary control, file organization and maintenance, and evaluation are covered later in this chapter.

It has become apparent that studies of user's needs are a very valuable tool to the system designer if he is to decide, with any accuracy, the needs for a particular information system and exactly what that system will be expected to do. Such men as Liston (79), Di Bello (28) of the Sea-Launched Information Analysis Center, and Dye, et al. (32) of the Air Force Dynamics Laboratory have illustrated how important user studies are in the design process. Saul and Mary Herner (50) have produced a review of the user studies employed by various organizations during the design of information retrieval systems. They point out seven rather surprising facts about these studies:

1) There have been relatively few techniques used.

2) There is diversity and ambiguity of language in the discussions of the techniques that have been used and their results.

3) There is a large diversity of users to which these few techniques have been applied.

4) There is a lack of innovation.

5) There is a failure to build on past gains.

6) There is a failure to profit from past mistakes.

7) There is frequently an absence of rigorous experimental designs.

The value of on-line retrieval systems was recognized a few years ago (Licklider, 77), but very few installations had the required time-shared computer facilities. At that time Kessler (64) and Salton (102) were able to use the M.I.T. facilities and did design on-line systems. Now that time-shared machines are more readily available, many on-line systems, such as CONVERSE (Drew, et al., 29), DOCUS (Informatics, 54), BOLD (Broko and Burnaugh, 11), MICRO (Smith, 110), AESOP (Spiegel, et al., 111), MOLDS (Allderige and Knoke, 1), and a system at I.B.M. Poughkeepsie (Holzbaur and Farris, 51) are being designed. Unfortunately, although descriptive reports on these systems are available, no literature is presented discussing the design methodology that was used in justifying the use of on-line versus batch-processing retrieval systems. Borko (10) lists five questions that need to be examined in some depth about on-line systems:

1) What user requirements can be better solved by an on-line system?

2) Can an on-line system use the same vocabulary organization and control techniques that manual and batch-processing storage and retrieval systems use?

3) Do on-line systems require a different file structure, and if so, how are these files to be organized?

4) Since on-line systems make use of interactive displays and data transmission facilities, what equipment should be used, and on what basis should this decision be made?

5) Does the evaluation of an on-line system require the use of techniques different from those appropriate for manual and batch-processing systems?

Subject Indexing, Indexing Terms, and Controlled Vocabularies

Basic to the problem of retrieving information from a store is the process of matching a stated need against some kind of label that indicates what information is contained within a document or other information carrying vehicle. Creating such labels involves analyzing the text containing the information and formatting the results of this analysis in an indexing language of some kind. The activity of generation and control of the indexing language chosen for a system is also necessary. Although there is no one text that completely covers the area of indexing and vocabularies, Kent's Textbook on Mechanized Information Retrieval (63), Machine Literature Searching by Perry et al. (96), and the three volumes of the Cranfeld Report by Cleverdon et al. (17, 18, 19) provide a sound basis from which a person interested in this area should begin his research. Some excellent reviews of the material available on indexing are provided by Sharp (107), Taulbee (119), and Baxendale (6). The theoretical concepts of indexing are covered by such men as Barnes (5) and Oliver et al. (92). Surveys of various types of indexing precepts and indexing languages appear in works by Coates (20) and Ullman (122).

A text by Perry and Kent (95), in which the use of links and roles in indexing is developed, discusses applications of indexing techniques. The performance of indexing systems is discussed by Cattley et al. (14) and Katter (60).

Vocabulary control is an integral part of indexing and indexing languages. There have been articles published concerning the definition and operation of an index language thesaurus by such people as Sharp (108) and Heald (47). The most informative papers on the effects of different types of vocabulary controls are included in a study by Williams (126). He discusses the five different approaches to vocabulary control:

1) Free-word vocabularies.
2) Selected or controlled vocabularies.
3) Generic vocabularies.
4) Coded vocabularies.
5) Classification and subject heading vocabularies.

Tanning (117) describes the effect of introducing a controlled vocabulary on the indexing and searching operations of a document retrieval system at the Air Force Materials Laboratory.

Most of the work done in the past few years on indexing and vocabularies for automated systems has been in the area of automated indexing. Lately there has been a disenchantment with these theories. Currently there seems to be some differance of opinion among the members of the research community. There are those who feel that automated indexing systems can be made operational (i.e., Lesk and

Salton, 74; and Minsky, 87). They agree, however, that the implementation of automated indexing systems will not be realized for some time. Exponents of the contrasting viewpoint include Dreyfus (30), Lipetz (78), Levien and Maron (76), and Jones (57).

Sharp (107) points out that some of the more important features emerging from current research tends to show that in small-to-medium systems, where no special conditions exist, over-elaboration of retrieval languages and their control devices cannot be justified. Similar findings have been made by others in researching small information retrieval systems. Cleverdon and Keen (19) found " . . . that the single term index languages are superior to any other type." Melton (86) observed the lack of effectiveness of structuring sentences in search files. Tague's (116) work indicated that association measures to extend a search prescription do not improve retrieval performance. Salton and Lesk (104) proposed the general thesis that, for high precision where low recall is not important, sophisticated techniques are superfluous. Blagden (8) concluded that links and roles serve no useful purpose in a large number of systems.

Search Files and Searching Mechanisms

File organization and search is concerned with the structure and operation of systems that store and retrieve large amounts of information. Basic concepts of file organization to facilitate information retrieval are discussed by Casey et al. (13) in Punched Cards: Their Applications to Science and Industry. The best text available describing computer file organization in detail is

Meadow's (85) The Analysis of Information Systems: A Programmer's Introduction to Information Retrieval. Other major works include A Theory for File Organization by Hayes (46) and Information Retrieval - A Critical Review edited by Schecter (106).

The measurement of association is central to the organization and search of machine files. Such measurement is needed in order to associate file records with each other and with search requests. This measurement is usually of the match-mismatch type: either a given file record matches the search specification completely or it is considered a mismatch. While this exact match association is being used by most systems, considerable experimental work with statistical measures of association is in progress. As a result of this work, associative measures have been developed sufficiently to be incorporated in some operating file systems. Jones and Curtice (55) have published a paper comparing the different term association measures. Salton (103, p. 621) provides experimental evidence of the usefulness of associative techniques:

> Specifically, the procedures based on synonym recognition, weighted content identifiers, cosine correlation to match documents and search requests, and document abstract processing are always more effective than methods using simple word stem matches (without synonym detection), nonweighted terms, correlation terms based only on the number of matching terms, and analysis procedures which consider only the titles of the documents being examined.

Research in this area is being continued by such people as Jones, et al. (56), Bryant, et al. (12), and Anderson, et al. (3).

Associated with file searching is the current research in natural language requests. Work is being performed on systems that can translate a user's natural language request into the structure required for the search system. Kellogg (61, 62) and Cautin (15) are working on these systems. Kellogg's system, CONVERSE, is being constructed at System Development Corporation under the time-sharing system implemented on the IBM Q-32. Easy English, as reported by Cautin, is being developed at the More School of Electrical Engineering.

The grouping of records within a search file is very important. To facilitate retrieval, it is desirable to group records that have the same or related content. Where the content of a record is identified by assigned index terms, a common method of grouping the records is to list them under each assigned index term (inverted file). Zunde, et al. (129) defined the distribution of index terms assigned to documents to maximize information transmission. From this, it was possible to compute the desirable number of records to which a term is assigned as a function of the average number of records assigned per term for the entire file.

Beyond the grouping of records that have individual index terms in common, it is possible to group records based on an evaluation of general similarity. Grouping records by a technique known as clumping is being studied at the Cambridge Language Research Unit. Jones and Jackson (58, 59) have reported on the current status of the unit's investigations. Ide, et al. (53), and Grauer and Messier (44) have reported on the current status of the clustering technique under investigation at Cornell. Another method of grouping is called the citation appraoch based upon the processing of "citing/cited by" relations that occur between records in a file. Levien and Maron (75) have developed this approach enabling a system to observe relations that exist between records in the file and to derive information not explicitly contained in the file.

Current literature concerned with file structure reflects two major research areas. One in the development of more general, and more complex, file structures. The other is the analysis of the effect of structure on access time. Most of the current structure development involves variations of a structure in which each logical record contains the information needed to generate the address of the next logical record linked to it. Such structures have been referred to as structured files, threaded-list files, multi-list files, etc. Gray (45) provides a valuable review of the many list-structure approaches that have been designed. A number of specific list-linking techniques have been published recently. Gabrini (39) reports on an application of multi-list files to the project TIP file of Physical Review articles. Ross (101) describes the AED free storage package used on the AED-1 compiler system.

Several recent papers have been concerned with the quantitative effect of the file structure on the expected average access time to obtain records from the file. Lowe (80) has written a doctoral dissertation in which he characterizes the average access time in terms of the number of index terms per physical record that can be obtained in an inverted file organization, as opposed to a linked-list structure.

Leimkuhler (73) takes a different approach. He considers the probability of use of the logical records and organizes the file into zones of roughly equivalent probabilities to minimize access time.

Systems Performance and Evaluation

At the 1968 American Documentation Institute convention, Murdock and Liston (89) pointed out that the most important measures of systems evaluation may be categorized by three measures: cost, performance, and benefits, and their interrelationships. Cost, although the most obvious measure to consider in evaluating a system, appears to be the hardest to obtain. Fasana (36) suggests that most " . . . librarians are frightened by costs associated with automation and, further, that many librarians do not know or do not want to know the costs of their present systems." Fasana's statement may be generally true, but there are a few exceptions. The ongoing evaluation of a large operating system, MEDLARS, has been reported by Cummings (26). Freeman (38) presented a cost analysis of various activities of the information systems used at Science Information Exchange. Several papers have been written on computer and programming costs that may be of value to a system designer and evaluator; they include articles by Erickson (34), Fasana and Fall (37), Marron and Snydelman (81), Nelson (91), and Weinwurm (123).

The principle measures of system performance include such things as speed, accuracy, quality, and extent of coverage. As King (66) points out, " . . . the single most controversial issue in system evaluation is the matter of relevance." Rees and Schultz (100) define relevance as a relationship between system responses and an expression of the user's request; on the other hand, Cuadra and Katter (25) define relevance as a relationship between a document and a user's information requirement statement. In a paper discussing the fundamental considerations of relevance in document selection, Perry (94) suggests that there are two types of relevance: "(i) relevance_1, i.e., the relevance of a selected document to a library client who may find it more or less useful for some purpose, in particular making a decision of any one or many kinds and (ii) relevance_2, i.e., the relevance of an index entry, classification heading or the like to the subject contents of a document." To clarify the concept of relevance, it becomes necessary to standardize the terminology. One common method, closely related to the two types of relevance presented by Perry, is to define the terms User Relevance Judgment and System Relevance Response. User Relevance Judgment (URJ) is the user's assessment of the degree of relevance; System Relevance Response (SRR) is the system's assessed degree of relevance. Clearly, one system objective is for SRR to correspond to URJ. A system's performance can be described by the relationship between URJ and SRR.

Even though the System Relevance Response may be multivalued, an operating system usually incorporates a decision point, Ds. Documents judged to be above the value of Ds in relevance are selected for retrieval; those below are not retrieved. Likewise, users apply a corresponding decision point, Du, where retrieved doucments either fall above this point and are considered relevant or fall below and are deemed unusable. Figure 3 incorporates URJ, SRR, Du, and Ds in describing what occurs when an information request is directed at a data base. The documents within the base will be separated into four catagories (X_{11}, X_{12}, X_{21}, and X_{22}). X_{11} represents the number of documents in the system considered relevant by both the System Relevance Response and the User Relevance Judgment, and so on.

		User Relevance Judgment		
		Relevant D_U	Not Relevant	Total
System Relevance Response	Relevant	X_{11}	X_{12}	$X_{1\cdot}$
D_S	Not Relevant	X_{21}	X_{22}	$X_{2\cdot}$
	Total	$X_{\cdot 1}$	$X_{\cdot 2}$	$X_{\cdot\cdot}$

Figure 3. Retrieval Profile with Dichotomous Values of User Relevance Judgments and System Relevance Response

Using a scheme as described in Figure 3, correlation measures may be used to describe accuracy. Pollock (97) discusses several correlation measure that may be applied, including rank order, Kendall's tau, Spearman's rho, and Salton's normalized measures. Searches are often made by inspecting documents sequentially by order of relevance response values. This technique can be visualized by starting the decision point Ds at the top of the System Relevance Response scale and lowering it to the bottom of the scale. Giuliano and Jones (40) assign SRR values by ranking documents in a prescribed manner. To measure search performance, they plot number of relevant documents retrieved

(X_{11}) against their ranked valued for all values of Ds. A similar operation was performed by Altmann (2) and King and Isakov (68) in which System Relevance Response values corresponded to sequential levels of breadth found by searching in a coordinate index system. The total number of documents retrieved ($X_{1.}$) was plotted against the recall ratio ($X_{11}/X_{.1}$) to obtain a gross cost to performance relationship. Cost was identified with total retrieval and performance with recall ratios. Wiederkehr (124) plots (a) the number of relevant documents retrieved (X_{11}) as a function of the number of documents retrieved ($X_{1.}$) and (b) recall ratio ($X_{11}/X_{.1}$) against fraction of documents retrieved by the system ($X_{1.}/X_{..}$). Swets (115) plots the distribution of the documents found to the left of Du (relevant documents) and the distribution of the documents found to the right of Du (non-relevant documents) against the relevance response scale as shown in Figure 4. The areas to the left of Ds are the probability of retrieving a document, given that it is relevant, $P(R/r)$, and the probability of retrieving a document, given that it is not relevant, $P(R/\bar{r})$. He then plots $P(R/r)$ versus $P(R/\bar{r})$ for various values of Ds and refers to this relationship as the operating characteristic curve that he sets forth as a measure of performance.

Recall ($X_{11}/X_{.1}$) and precision ($X_{11}/X_{1.}$) ratios are the primary accuracy measures currently being used. These measures have come under considerable criticism by such men as Fairthorne (35), McCarn and Stein (83), and Tukey (121). King suggests that recall and precision ratios should be used primarily to highlight the fact that a system is not

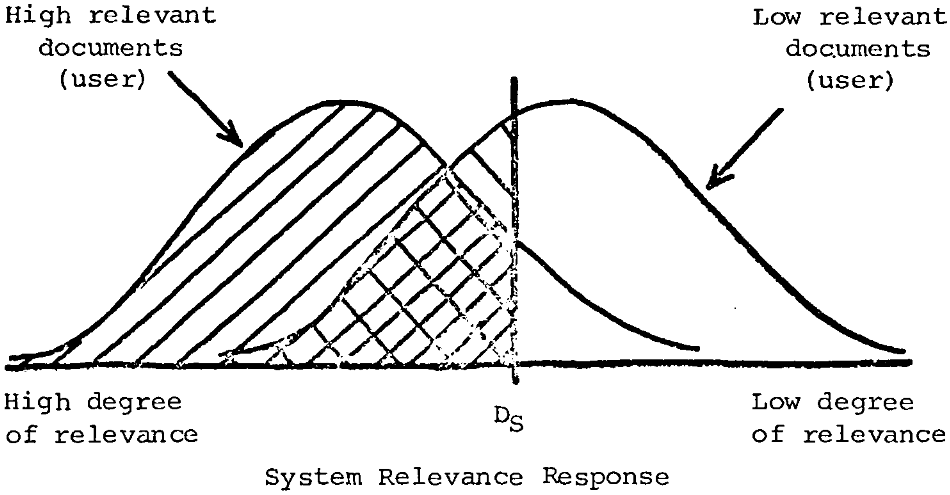

Figure 4. Distribution of High Relevant Documents (to the Left of D_S) and Low Relevant Documents (to the Right of D_S) from User Relevance Judgments Plotted Against the System Relevance Response Scale

performing satisfactorily. When it is determined that a system is not operating satisfactory, further analysis must be performed to determine why. Studies by Hooper (52), King (67), Lancaster (71), McCarn and Stein (83), and Taulbee (118) emphasize failure analysis to determine the extent to which failures (X_{21} and X_{12}) are attributable to incorrect query formulation, indexing errors, mechanical errors, incorrect screening, etc.

A measure of the benefit of a system is much more difficult to define than its cost or performance. Dugger and Klinger (31) attempted to determine (by questionnaire) the value of information received from a system in terms of dollars or man-hours saved. Emery (33) suggests that information has value to the extent that it contributes to the achievement of an organization's goals. He defines "utility" as a single index of an organization's well-being and notes that information must lead to an increase in utility. He points out that this value depends on system performance. Goffman and Newill (41) also propose a utility function on an empirical basis. A utility function based on theoretical grounds has been developed by Good (42).

CHAPTER III

DEVELOPMENT OF THE CANED INFORMATION RETRIEVAL SYSTEM

The CANED System

Under National Science Foundation Grant No. GN-735, the Office of Arid Lands Studies of The University of Arizona has undertaken to develop and implement an information storage and retrieval system to be applied to its substantial arid-lands bibliographical resource, as well as to provide for the future growth of this resource. In partial fulfillment of this objective, the CANED Information Retrieval System has been developed and is now in limited experimental use by the Office of Arid Lands Studies. It is a high-speed, random access system utilizing the University of Arizona's Control Data Corporation 6400 Computer.

The CANED System stores bibliographical information (author, date of publication, publisher, and title), abstract, and key-words for each document in the arid-lands data base. Retrieval is accomplished on four categories: author, date of publication, publisher, and key words. The use of the Boolean Functions, "AND," "OR," and "NOT," allow for combinations of these four categories to be used in making up requests for information. In conjunction with the information retrieval system, a cross-reference index of key words is provided. This index provides a listing of all key words and their associated terms that have to do with arid lands material.

The Initial Study

The first step in the design process of the CANED system was a meeting with Miss Patricia Paylore, Research Associate with the Office of Arid Lands Studies. The purpose of this meeting was to establish (a) why an information retrieval system was needed, (b) what it was expected to do, and (c) what problems were to be solved with such a system. She explained that the only existing bibliographic data base of arid lands information was compiled by her office under U.S. Army Natick Laboratories contract DA49-092-aro-71. The material contained in this data base covers the fields of vegetation, fauna, geomorphology and surface hydrology, surface materials, weather and climate, and regional/cultural features. While many of the references within this base appear in existing indexing and abstracting tools, equally as many do not. Furthermore, retrieval of those references that do appear in the indexing and abstracting services involves lengthy and often futile searches. This is due partly to the non-standard key word indexing and subject terminology applied and partly to the fact that available abstracts and/or evaluations of documents rarely focus on applications to or for aridity. Thus, the CANED system was needed to make available a collection of bibliographic information to scientists working in the interdisciplinary field of arid lands studies. The proposed system would be expected to provide the translation of the bibliographic data base compiled by the Office of Arid Lands Studies into a machine readable form and to provide a method of retrieving citations from this data base by subject matter. It would aid the research community in narrowing the scientist's search for information by defining descriptive indexing terms more exactly and by broadening the range of citations available by the greater scope of material being indexed.

The next step was to define the three basic design criteria: effectiveness, resource, and time. The effectiveness or benefit criteria included establishing a data base of arid lands material which would be available to all interested parties and providing an effective, low cost method of subject search on the data base to retrieve documents relevant to a user's request. Evaluation of resource criteria required that the available resources be taken into consideration. These initial resources consisted of (a) the bibliographic data base compiled by the Office of Arid Lands Studies, (b) funds available from a National Science Foundation grant (GN-735) to initially establish the information system, and (c) the University of Arizona's Control Data 6400 computer. The schedule criteria required that a pilot information retrieval program be operating within six months and that the retrieval system be operating on a limited scale within twelve months. Therefore, the initial project was to establish a machine stored data base with appropriate retrieval methods within a twelve month period. The only environmental constraint was that the resources available during the design and implementation of this pilot system would be from the Office of Arid Lands Studies of The University of Arizona.

After the design criteria had been defined, the probable users of the system were consulted. The Office of Arid Lands Studies identified various user groups who would have need of such a data base. Due to the interdisciplinary type of materials involved, the users of the system come from many widespread fields: agronomists, anthropologists, architects, astronomers, atmospheric physicists, botanists, climatologists, demographers, doctors, ecologists, economists, engineers, entomologists, ethnologists, geneticists, geographers, geologists, horticulturists, hydrologists, microbiologists, nutritionists, paleontologists, plant pathologists, radiation biologists, sociologists, soil scientists, space scientists, zoologists, as well as travelers, students, teachers and government officials. While all of these people would have use for the data base, the primary user group would consist of the Office of Arid Lands Studies at The University of Arizona and the other five arid lands research centers in the western United States: The Dry Lands Institute at The University of California at Riverside, the Desert Resources Institute at The University of Nevada, The International Center for Arid and Semi Arid Lands Studies at Texas Technical University, New Mexico State College at Los Cruses, and Utah State College at Logan. It would be impossible at the onset of this project to discuss system requirements with the bulk of this user population. Therefore, it was decided to initially rely on the Office of Arid Lands Studies as the primary user group.

Material for the user study was drawn from the Arid Lands Information Service of the Office of Arid Lands Studies. This service is intimately familiar with the actual working language of the user population and has considerable experience with requests for information about arid land subjects. From discussions with this information service, four important points emerged about the user population:

1) The bulk of the information requests concern general topics rather than specific technical points.

2) Many users of the service come from the less technically developed areas such as Sudan, Pakistan, Chile, etc.

3) The terminology of most requests are jumbled and often contradictory.

4) Requests for information come primarily from the research community rather than from applied scientists.

After the system criteria and constraints had been defined, and the users' needs had been discussed, the components of the proposed pilot system were evaluated, organized, and set forth in the following outline:

A) PURPOSE COMPONENTS

1) Type of Users

a) The initial user population consists of the Office of Arid Lands Studies at The University of Arizona.

b) Members of the user population are primarily research personnel involved in projects associated with arid land areas.

c) The population is restricted to the 40 to 50 people already acquainted with the information service of the Office of Arid Lands Studies.

d) User needs consist of literature searches to provide bibliographic information on material related to particular subject areas.

2) Subject Discipline

Subjects covered within the data base will be those initially used by the Office of Arid Lands Studies to compile the bibliographic information for the U.S. Army Natick Laboratories.

3) File Size

The initial file size will be approximately 20,000 documents (those used for the Natick Laboratory study).

B) FUNCTION COMPONENTS

1) Acquisition

The acquisition of material for the data base will be handled by the Office of Arid Lands Studies. (Acquisition sources, policies, and procedures are extant, having been established for the Natick Laboratories project.) It should be noted that material in the data base need not be the most recent and advanced material available. A large portion of the user population is comprised of underdeveloped countries that are interested in the more 'classic' documents, those that pertain more to the current 'state of the art' that exists within their countries. Thus, older material will be considered for the system as well as the most recent publications.

2) Indexing Language

The source of the indexing language will be a Thesaurus of Arid Lands Terminology. Vocabulary control will be under the direction of Miss Paylore of the Office of Arid Lands Studies who is extremely qualified as a librarian and an authority on arid lands terminology. The vocabulary size has been established at approximately 10,000 terms for the existing data base of 20,000 documents.

The sources of material for indexing consist of approximately 65% full texts, 20% prepared abstracts, and 15% bibliographic information alone. Initially, all indexing will be done by Miss Paylore. She will be responsible for producing such information as indexing instructions, policies, and manuals. It is extremely important that the person responsible for indexing is experienced both in arid lands terminology and in the operations of automated retrieval systems.

3) Coding

Coding instructions will be provided in the system operating manuals. The primary consideration in the development of coding policies, is to keep them as simple and straightforward as possible. Coding personnel will not be required to have any special knowledge of library science beyond that expected of any secretary.

4) File Organization

Organization of the files will be handled by the computer programs. Retrieval will be accomplished using uniterms organized in an inverted file structure. Due to its initial size, no structuring of the data base will be necessary.

The only handling of materials will involve the delivery of input and retrieval of output from the Computer Center.

5) Question Analysis

Users of the system will correspond with the Office of Arid Lands Studies either by letter, wire, phone, or in person. The personnel at the office will translate the user's request into machine compatable terminology using the compiled thesaurus as a guide. When any ambiguities arise in an information request, the office will immediately discuss the problem directly with the user, either by telephone or in person.

Two types of queries will be available: a general query used for broad based subject area searches and a very precise query for search on specific material within a very narrow field.

6) Searching Procedures

Boolean logic (AND, OR, NOT) will be used in search formulation. Search parameters will consist of Subject, Author, Publisher, and Date of Publication. The search scope may be adjusted via the generality or specificity of terms used.

7) Dissemination

Output from a search of the data base will consist of bibliographic information, abstract, and key word descriptors of all documents that satisfy the search request. Output will be on standard 11" x 15" computer paper as reproduction of the output should not be needed. Dissemination will be handled by the most convenient method available. Once a user's request for information has been answered, the user will be asked to return an evaluation of the system, in general, and of the materials received.

Initial Design

After the components of the proposed system had been considered, the first guess solution, or pilot system was constructed. By placing the emphasis on the human element, it was felt that a simple, efficient, and inexpensive automated information system, with performance equal to the more sophisticated systems, could be implemented. This supposition led to close consideration of the trade-off of man versus machine functions throughout the design stages.

The Cranfield Reports (Cleverdon, et. al.; 17, 18, 19) tend to illustrate that the effectiveness of an information retrieval system is not due to the method of retrieval employed by the system, but rather to the method of indexing and query formulation. Therefore, to prevent many of the inconsistencies encountered by other systems, the personnel of the Office of Arid Lands Studies are responsible for both the indexing of material and the formulation of user requests into system compatible queries. To facilitate uniformity in the indexing of documents, a thesaurus is provided that lists all valid arid land terms. The thesaurus is constructed so that all the related, broader, and narrower terms are listed for each entry. If a user has only vague concepts of what information he wants from the system, the thesaurus is also useful in providing a method of formulating a more exact query.

As each new document is reviewed by an indexer, two levels of keyword descriptors are supplied. The first level contains only general keywords describing the general content of the article. The second level contains a refinement of the initial selection of descriptors. This makes it possible to retrieve the document when a query contains either general or specific descriptors about particular subject matter covered within that document.

A list is provided by the computer of all descriptors currently in the data base. At the time a query is formulated, but before it goes to the computer, it is checked against this list. This prevents machine time being wasted in determining that a query contains a descriptor in any way incompatible with the system.

The coding required to communicate with the system has been kept very simple. Once a query has been formulated, or a document marked for addition to or deletion from the data base, any secretary within the office can easily code and keypunch the required information and see that the proper computer programs are executed.

Saul Herner (48) has found that a user population consisting primarily of applied scientists is more interested in automated information retrieval systems than a population of pure scientists. Pure scientists prefer to browse through a library section pertaining to their current interests rather than to pose direct questions to a computer. Therefore, this system, oriented primarily to the pure scientist, is designed to store citation entries (bibliographic information, abstract, and key word descriptors). By producing machine searches with high recall, the requester may "browse" through the retrieved abstracts, gleaning more information from them than from titles alone.

In adapting the system for the use of any small library, changes would not be necessary in the computer programs. The primary obstacles would be encountered in achieving efficient retrieval. This objective would be met by establishing an appropriate controlled vocabulary, using proper indexing techniques, and gaining experience in query formulation.

The retrieval system is of the inverted file type used with random access disk storage. The method of indexing is uniterm. Separate magnetic tapes store the large tables and the data base. There is no need to keep this information on the disk (on-line) since users will not have direct access to the computer. Were a real-time computer system available, however, the costs incurred would probably be prohibitive for the amount of expected use of the system. Thus, queries are batch processed. When the retrieval program is to be used, the large tables are loaded onto the disk. Using the input queries and the disk tables, the program compiles a list of the reference numbers of documents satisfying the queries. After all queries have been processed, this list of document numbers is ordered and one pass is made through the citation tape retrieving the actual citation entries.

To provide a useable system, four programs in addition to the retrieval program were necessary. These programs provide the capability to (a) add documents to the system, (b) drop documents from the system,

(c) list documents in the data base that have been infrequently retrieved, and (d) print out the tables and the data base in the event of a system malfunction. A frequency table is kept by the system containing the number of times each citation in the data base has been retrieved. This allows for periodic inspection of the infrequently retrieved documents to insure that they have been entered into the system correctly and/or that they belong in the data base. When it is desired to check the infrequently retrieved documents, the print program is executed.

Input to the program consists of a number (the frequency number) which calls for the program to print out all citations in the data base that have been retrieved that number of times or less. The print program also provides the option of zeroing or erasing the frequency table.

Procedures for the operation of the information retrieval, document addition, and document delition programs have been established and are given in figures 5, 6, and 7 below.

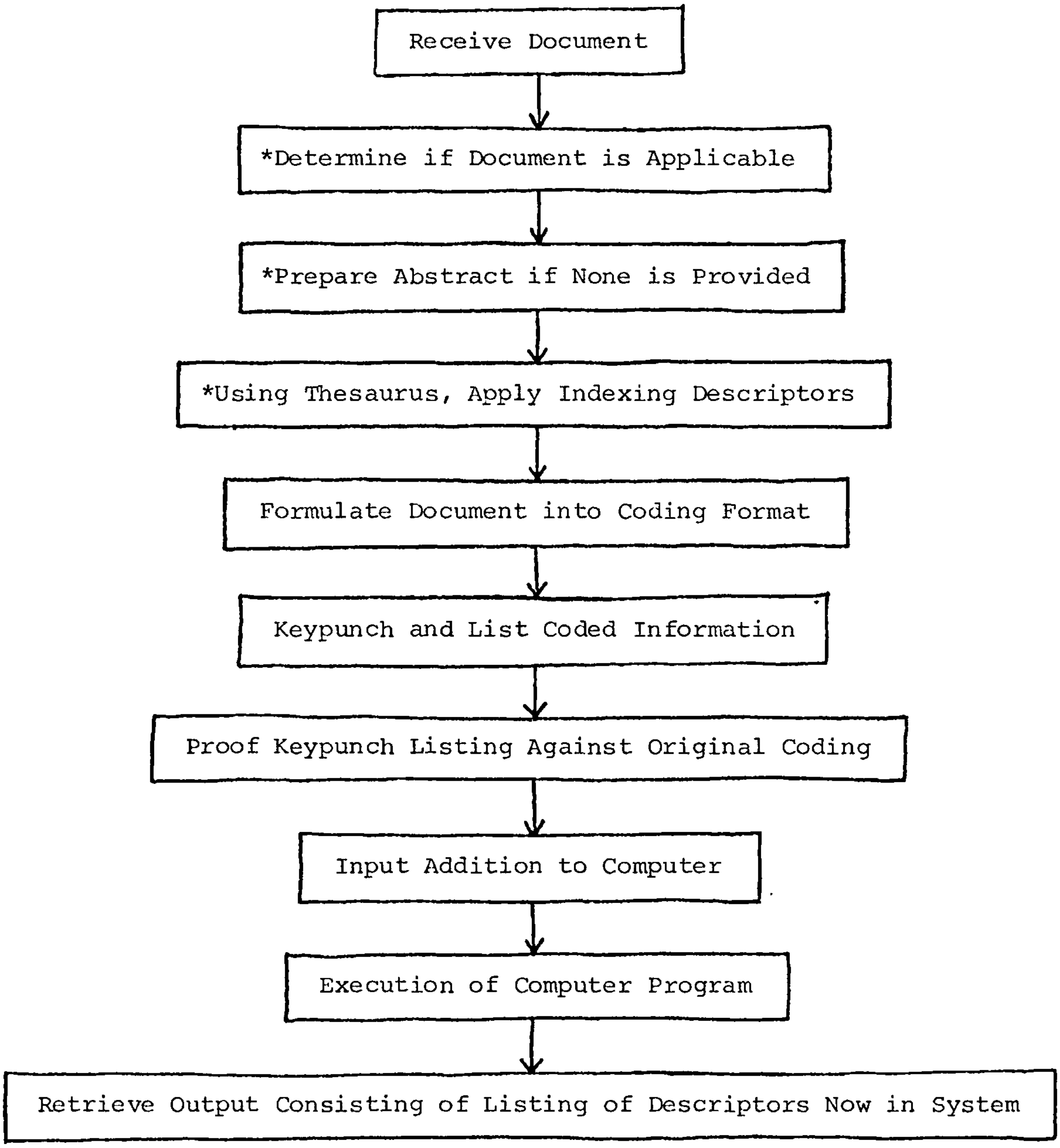

Figure 5: Procedure for Adding Documents to the System

*These operations are to be handled by Miss Paylore. All others may be handled by any employee of the Office of Arid Lands Studies.

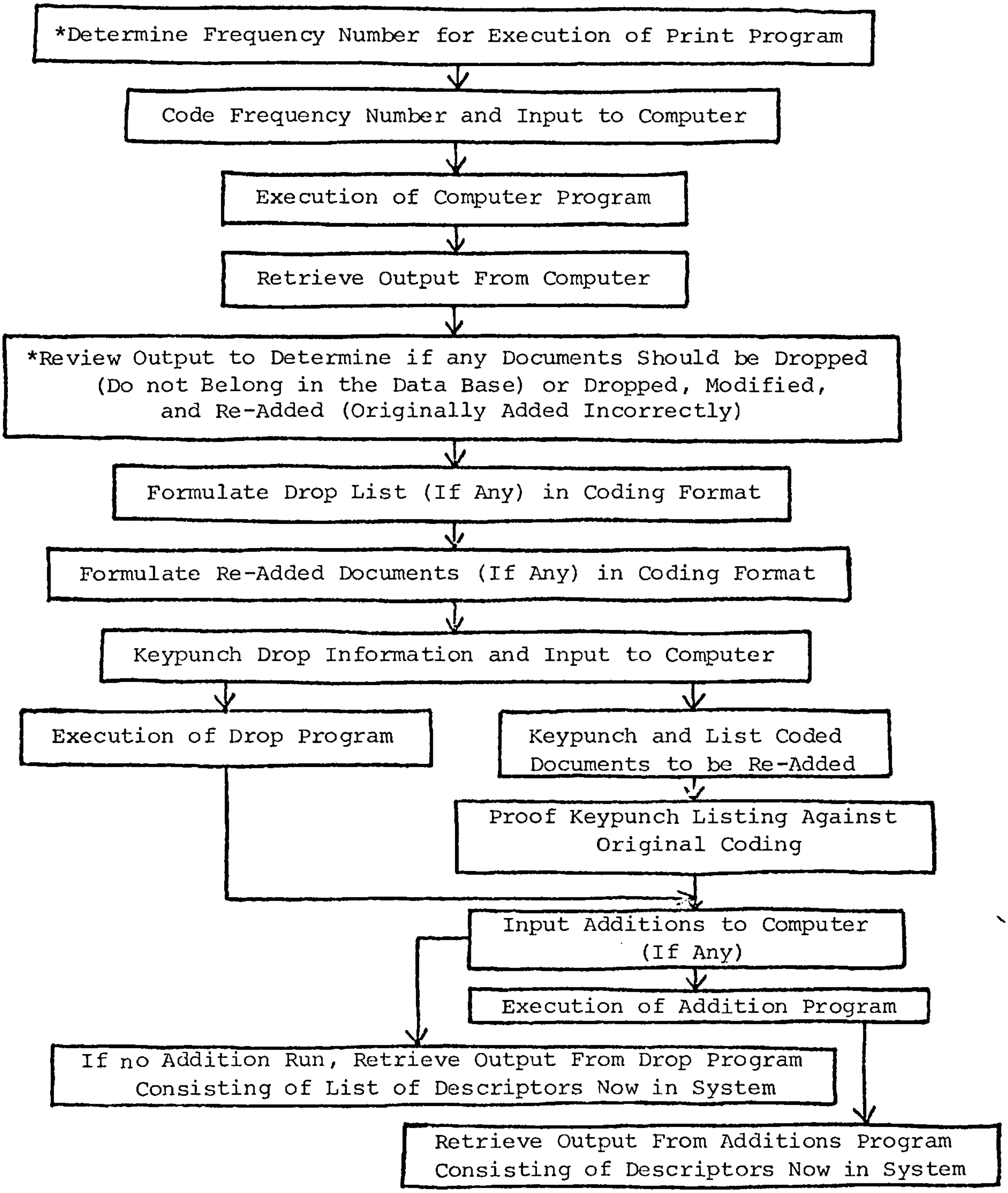

Figure 6: Procedure for Deleting Documents from the System

*These operations are to be handled by Miss Paylore. All others may be handled by any employee of the Office of Arid Lands Studies.

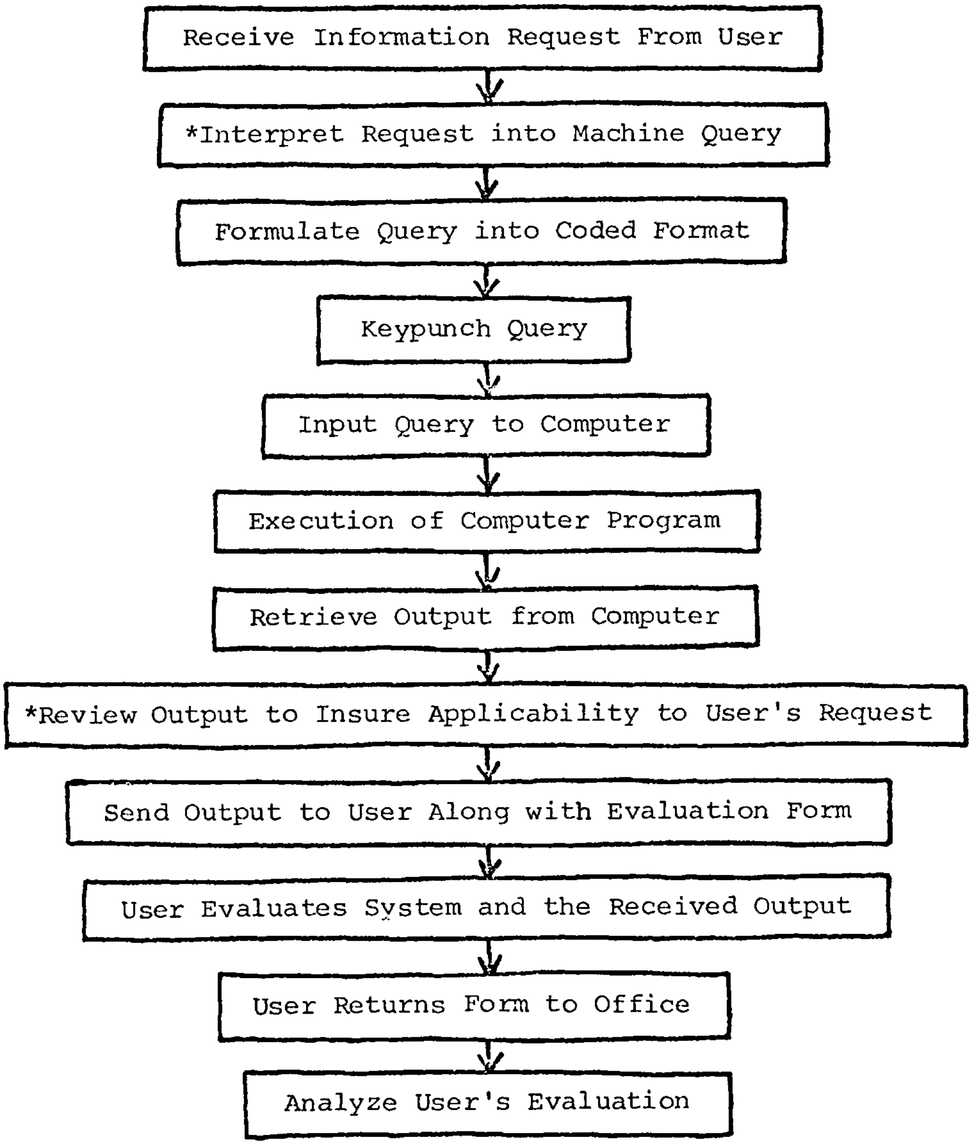

Figure 7: Procedure for the Retrieval of Information from the System

*These operations are to be handled by Miss Paylore. All others may be handled by any employee of the Office of Arid Lands Studies.

The retrieval system programs, along with the thesaurus program, have been written and are now in operation. The flowcharts of the computer programs are provided in figures 8 and 9.

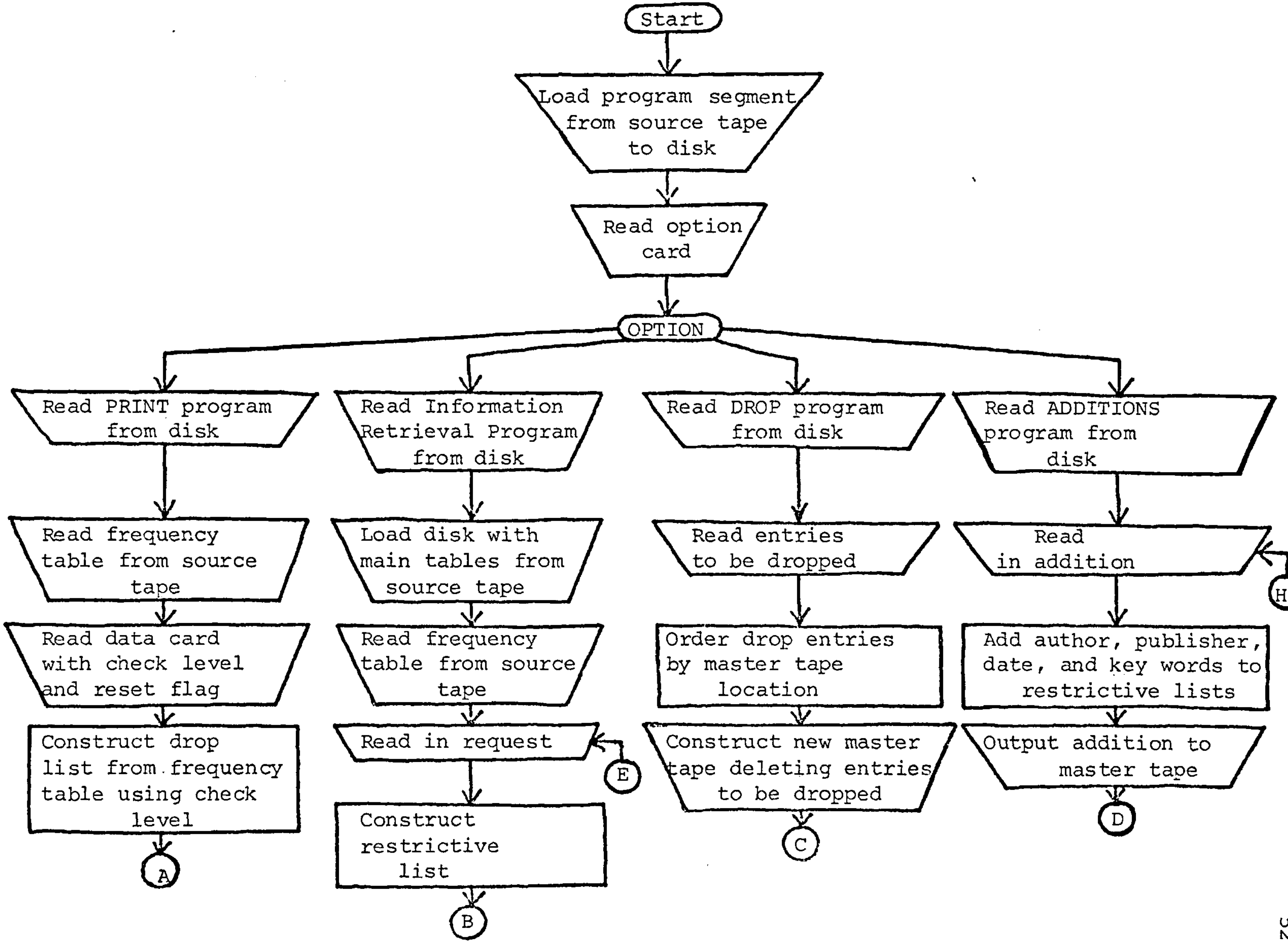

Figure 8: Flowchart of the Retrieval System Programs

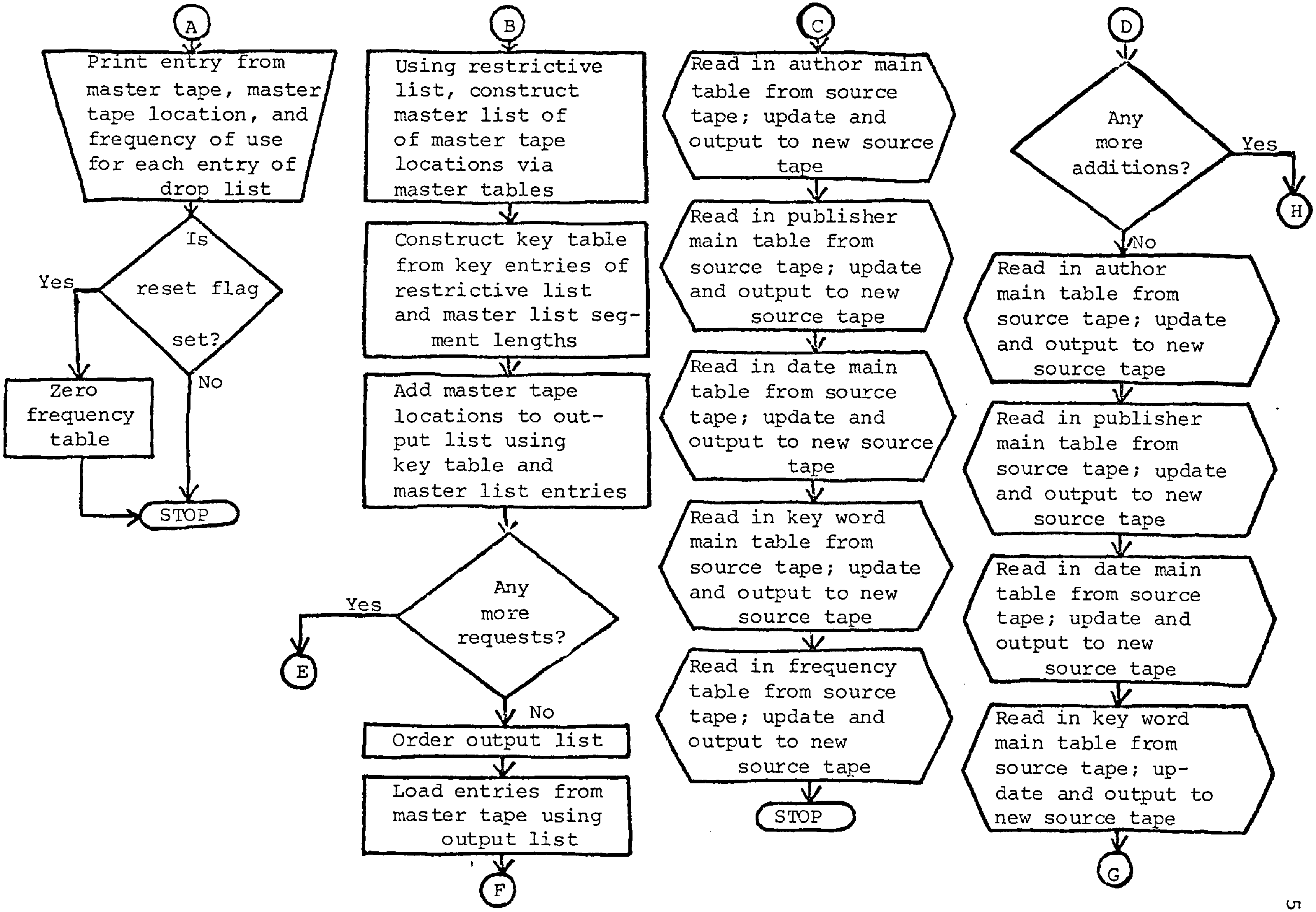

Figure 8: Flowchart of the Retrieval System Programs (Continued)

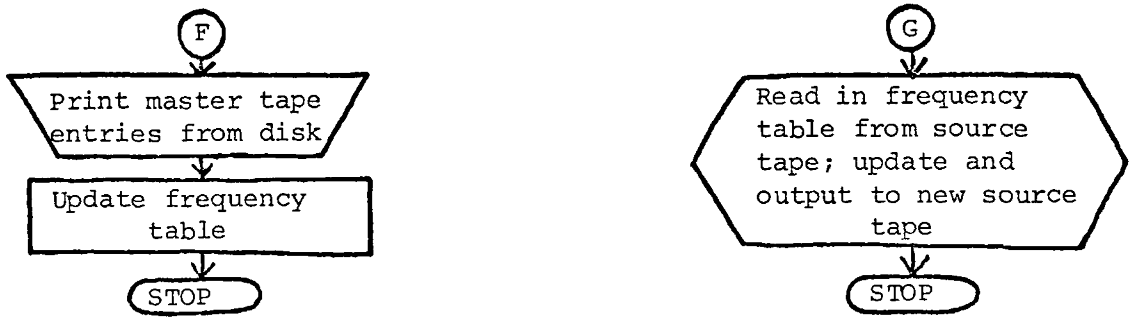

Figure 8: Flowchart of the Retrieval System Programs (Continued)

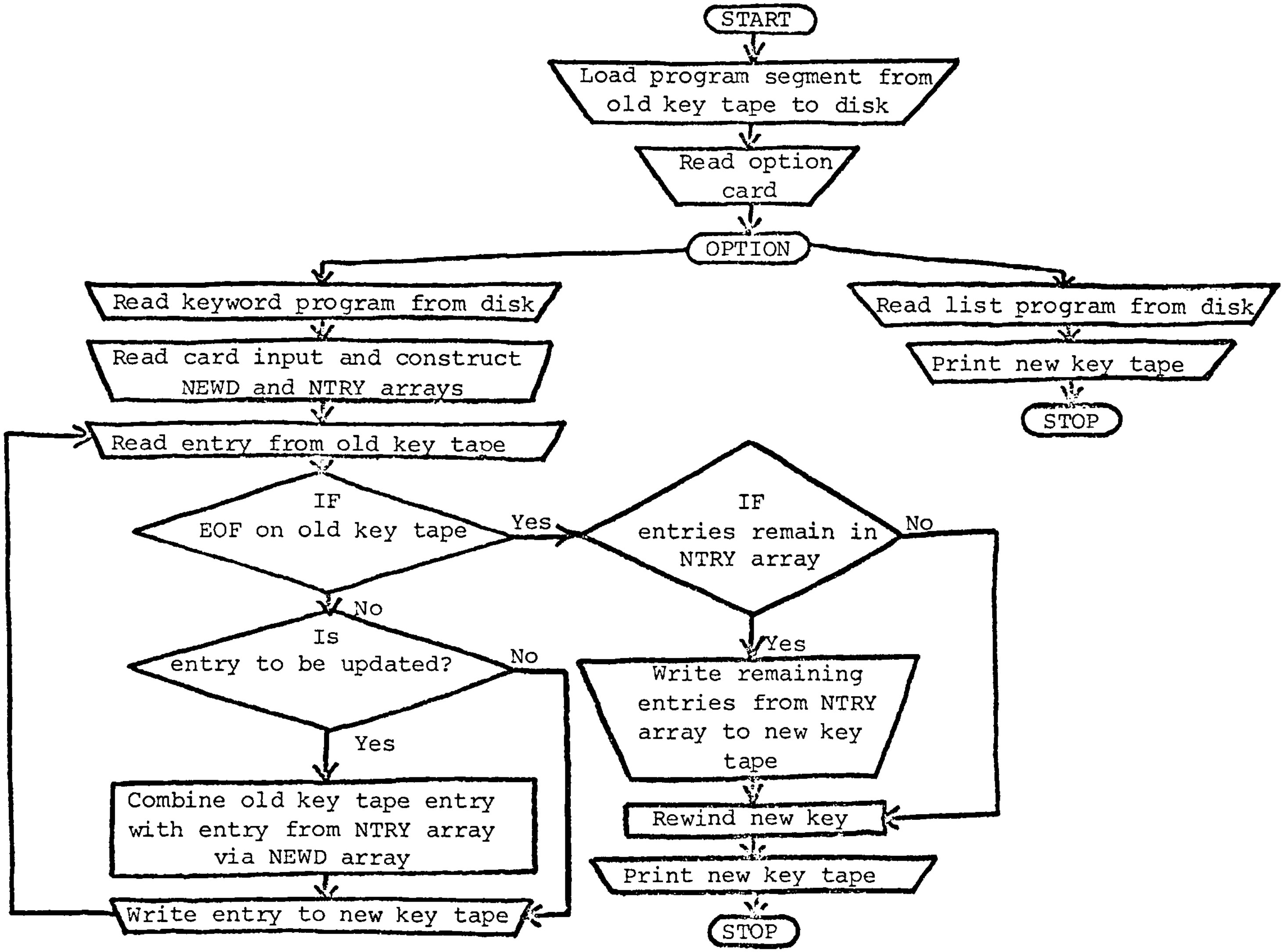

Figure 9: Flowchart of the Thesaurus Programs

Currently, a data base of approximately 500 test documents and the thesaurus are being loaded into the machine. The next step will be evaluation of the operating system. Such things as system performance, operating procedures, and retrieval costs need to be carefully evaluated. As soon as the system appears to be operating satisfactorily, the remaining documents from the Arid Lands Office's bibliographic resource will be loaded into the retrieval system, completing the initial design phase of the information system.

CHAPTER IV

THE FUTURE OF THE CANED INFORMATION RETRIEVAL SYSTEM

Phase I of the development of the CANED system was described in Chapter III. As this phase approaches completion, the expansion of the system into a cooperative enterprise including the other five arid lands research centers seems eminent. Each center, specializing in a different facet of arid lands studies, will be responsible for contributing indexed articles concerned with its particular specialized area. Accommodation of the proposed increase in contributors will require major revisions and a complete re-evaluation of the system. This will be designated as Phase 2 and will be approached by again using the design procedures as set forth in Chapter I.

The increased scope of the proposed system will necessitate a meeting of key personnel from the six arid lands research centers. The objectives of this meeting will be twofold: (a) to establish cooperation among the six centers and to set up appropriate communication channels and (b) to determine what kind of an information system is desired by the centers. Such questions as "What do you expect the system to do?", "What do you want from the system?", and "What do you have to contribute to the system?" will be considered.

Each cooperating center will provide a detailed user study. The combination of these six studies should provide a rather comprehensive picture of the information system's user population. The designing and assembling of the processes and procedures comprising the system will involve the assignment of system responsibilities and duties to the different offices. It is critical during this time that communication channels be effectively used to insure complete cooperation among the centers.

Once the initial design is established and approved, a pilot system will be constructed. Before the system is considered operational, completing Phase 2, it will undergo extensive testing and evaluation.

Although the immediate future of the CANED system consists of the Phase 2 design described above, there will continue to be modifications after the system becomes operational. The type and extent of the changes that will take place depend on three basic factors: (a) the system's operational expenses, (b) the users' needs, and (c) the growth of the data base.

The National Science Foundation is funding the project through the initial design phases. However, the operational system is expected to be self-supporting and any changes will have to be made within the constraint that operational costs be met.

As the system becomes more widely recognized by the user public, their demands on it will change. As the users become more sophisticated in their use of the system, the system will have to become correspondingly more sophisticated. Although the system will remain as it has been initially designed, a research service that performs literature searches upon request, the system may be asked in the future to supply other services. Magnetic tapes of the data base may be requested by users desiring to incorporate arid lands materials into their own information systems. There may be a need for published materials such as a KWIC index or a manual of the abstracts in the data base along with subject and author indexes. A selective dissemination of information service may be incorporated as part of the system. The fact that the CANED system will be the only comprehensive data base of arid lands materials available insures that there will be a wide variety of uses for it in the future.

The retrieval programs in the system will have to be continually re-evaluated and modified. As the data base grows, more sophisticated retrieval devices will become effective. Two such devices, the structuring of the files within the data base and the addition of links and roles to the retrieval programs, may become useful when the base reaches sufficient size.

Thus, to insure that the system remains an effective and useable tool, the managers of the system will be expected to (a) continuously evaluate and modify the operating system, (b) remain sensitive to changes in user needs, (c) keep abreast of the 'state of the art' in information science, and (d) continually explore the marketing potentials and possibilities of the system.

In conclusion, then, a basic principle of the continued effective operation of an information system is the "evaluation and modification cycle." As mentioned before, the most important consideration in the effective design, evaluation, or modification of a system is that the system be considered as a whole. Each element of a system must be planned, designed, and evaluated by explicitly considering its interrelationships with the other elements of the system and by optimizing its cost and benefit as part of the total system's cost and benefit.

The applications of systems engineering presented in this paper are by no means complete. Some basic systems engineering procedures have been discussed and a preliminary model has been established. There are immediate opportunities to apply systems engineering methods in several areas of the development of the CANED system. For instance, systems engineering should be applied to: (a) the evaluation of the performance of the system, (b) the development of the expanded system, incorporating the six arid lands research centers, (c) the establishment of cost and benefit criteria and the optimization of the system with respect to these criteria, and (d) the establishment of a method for determining the costs of the system and incorporating effective user billing procedures.

SELECTED BIBLIOGRAPHY

1. Allderige, J. M.; Knoke, P. J. Design Considerations for an On-Line Management System. Syracuse University Research Corporation, New York, July 1966.

2. Altmann, Berthoud. "A Multiple Testing of the National Language Storage and Retrieval ABC Method: Preliminary Analysis of Test Results." American Documentation, 18: 33-45, January 1967.

3. Anderson, Ronald R.; Kasarda, A. J.; Reed, D. M. Experimental Retrieval System Studies, Report No. 3. Center for the Information Sciences, Lehigh University, Bethlehem, Pa., April 15, 1967.

4. Artandi, Susan. An Introduction to Computers in Information Science. Scarecrow Press, Inc., Uchen, New Jersey, 1968.

5. Barnes, Robert F. Mathematico-Logical Foundations of Retrieval Theory: General Concepts and Methods. Center for the Information Sciences, Lehigh University, Bethlehem, Pa., November 1965.

6. Baxendale, Phyllis. "Content Analysis, Specification, and Control," In: Annual Review of Information Science and Technology, Volume 1. John Wiley and Sons, New York, 1966.

7. Becker, Joseph; Hayes, Robert M. Information Storage and Retrieval: Tools, Elements, Theories. John Wiley and Sons, New York, 1963.

8. Blagden, J. B. "How Much Noise in a Role-Free and Link-Free Co-Ordinate Indexing System?" Journal of Documentation, 22:203-209, September 1966.

9. Borko, Harold. The Analysis and Design of Information Systems. System Development Corp., Santa Monica, California, November 10, 1966.

10. Borko, Harold. "Design of Information Systems and Services," In: Annual Review of Information Science and Technology, Volume 2. John Wiley and Sons, New York, 1967.

11. Borko, Harold; Burnaugh, H. P. "Interactive Displays for Document Retrieval." Information Display, 3: 47-90, September 1966.

12. Bryant, Edward C.; Searls, D. T.; Shvmway, R. H.; Weinman, D. G. D. G. Associative Adjustments to Reduce Errors in Document Screening. Westat Research, Inc., Bethesda, Maryland, March 31, 1967.

13. Casey, Robert S.; Perry, James W.; Berry, Madeline M.; Kent, Allen. Punched Cards, Their Applications to Science and Industry, Second Edition. Reinhold Publishing Company, New York, 1958.

14. Cattley, J. M.; Moore, J. E.; Banks, D. G.; O'Leary, P. T. "Inter-Index Patent Searching by Computer." Journal of Chemical Documentation, February 6, 1966.

15. Cautin, Harvey; Lowe, T. C.; Rapp, F. A.; Rubinoff, M. An Experimental On-Line Information Retrieval System. University of Pennsylvania, Moore School of Electrical Engineering, Philadelphia, April 1967.

16. Chapanis, Alphonse. "On the Allocation of Functions Between Men and Machines." Occupational Psychology, 39: 1-11, January 1965.

17. Cleverdon, Cyril; Mills, Jack; Keen, Michael. Factors Determining the Performance of Indexing Systems, Volume 1: Design, Part 1: Text. Aslib Cranfield Research Project, Cranfield, Bedford, England, 1966.

18. Cleverdon, Cyril; Mills, Jack; Keen, Michael. Factors Determining the Performance of Indexing Systems, Volume 1: Design, Part 2: Appendices. Aslib Cranfield Research Project, Cranfield, Bedford, England, 1966.

19. Cleverdon, Cyril; Mills, Jack; Keen, Michael. Factors Determining the Performance of Indexing Systems, Volume 2: Test Results. Aslib Cranfield Research Project, Cranfield, Bedford, England, 1966.

20. Coates, E. J. "Scientific and Technical Indexing." Indexer, 5: 27-34, Spring 1966.

21. Corrigan, Robert E.; Kaufman, Roger A. Why Systems Engineering. Fearon Publishers, Palo Alto, California, 1966.

22. Cuadra, Carlos A. Annual Review of Information Science and Technology, Volume 1. John Wiley and Sons, New York, 1966.

23. Cuadra, Carlos A. Annual Review of Information Science and Technology, Volume 2. John Wiley and Sons, New York, 1967.

24. Cuadra, Carlos A. Annual Review of Information Science and Technology, Volume 3. Encyclopedia Britannica, Inc., Chicago, 1968.

25. Cuadra, Carlos A.; Katter, R. V. "The Relevance of Relevance Assessment," In: Proceedings of the 30th Annual Meeting of the American Documentation Institute, New York, October 1967.

26. Cummings, Martin M. "Needs of the Health Sciences," In: Electronic Handling of Information: Testing and Evaluation. Academic Press, London, 1967.

27. Dearden, J.; McFarlan, F. W. Management Information Systems: Text and Cases. University of Texas Press, Austin, 1966.

28. DiBello, Franz A.; Viil, Heino. The Definition of Requirements for a Sealaunched Missile Information-Analysis Center. Vitro Laboratories, Silver Springs, Maryland, November 1965.

29. Drew, D. L.; Summit, R. K.; Tanara, R. I.; Whiteley, R. B. "An On-Line Technical Library Reference Retrieval System." American Documentation, 17: 3-7, January 1966.

30. Dreyfus, Herbert L. Alchemy and Artificial Intelligence. Rand Corporation, Santa Monica, California, December 1965.

31. Dugger, Edward; Klinger, R. F. "User Evaluation of Information Services," In: Information Retrieval: The User's Viewpoint, An Aid to Design. International Information, Inc., Philadelphia, Pa., 1967.

32. Dye, H. E.; Bloomquist, C. E.; Wilmot, C. H. Flight Control Data and Information Exchange Program, Part 1: Investigation of Need and Feasibility. Report for June, 1964 - April, 1965, Systems Technology, Inc., Inglewood, California, November 1965.

33. Emery, James. "Economics of Information." Presented at International Systems Meeting, Detroit, Michigan, October 1967.

34. Erikson, Warren J. An Analytical Cost Comparison of Computer Operating Systems, System Development Corp., Santa Monica, California, June 30, 1967.

35. Fairthorne, Robert A. Unification of Theory and Empiricism in Information Retrieval, Final Scientific Report, Herner and Company, Washington, D. C., October 5, 1967.

36. Fasana, Paul J. "Determining the Cost of Library Automation." American Library Association Bulletin, 61: 656-661, June 1967.

37. Fasana, Paul J.; Fall, J. E. "Processing Costs for Science Monographs in the Columbia University Libraries." Library Resources and Technical Services, 11: 97-114, Winter 1967.

38. Freeman, Monroe E. "Determining Costs of Information Systems." Journal of Chemical Documentation, 7: 101-106, May 1967.

39. Gabrini, P. J. Automatic Introduction of Information into a Remote-Access System: A Physics Library Catalog. University of Pennsylvania, Moore School of Electrical Engineering, Philadelphia, November 1, 1966.

40. Giuliano, Vincent E.; Jones, P. E. Study and Test of a Methodology for Laboratory Evaluation of Message Retrieval Systems, Interim Report. Decision Sciences Lab., Electronic Systems Division, Air Force Systems Command, U.S. Air Force, L. G. Hanscom Field, Bedford, Massachusetts, August 1966.

41. Goffman, N.; Newill, V. A. "A Methodology for Test and Evaluation of Information Retrieval Systems." Information Storage and Retrieval, 3: 19-25, August 1967.

42. Good, F. J. "The Decision-Theory Approach to the Evaluation of Information-Retrieval Systems." Information Storage and Retrieval, 3: 31-34, April 1967.

43. Goodman, A. F.; Gainen, L.; Beum, C. O. Jr. Complete System Analysis: Quantitative System Analysis, Computer Stimulation, and System Optimization. Douglas Aircraft Co. [Santa Monica, California], March 1967. [Douglas Paper No. DP-4431].

44. Grauer, Robert T.; Messier, M. "An Evaluation of Rocchio's Clustering Algorithm." Information Storage and Retrieval, June 1967.

45. Gray, J. C. "Compound Data Structure for Computer Aided Design: A Survey," In: Proceedings of the 22nd National Conference, Association for Computing Machinery, Thompson, Washington, D. C., 1967.

46. Hayes, Robert N. "A Theory for File Organization," In: On-Line Computing: Time-Shared Computer Systems, McGraw-Hill, New York, 1967.

47. Heald, J. Heston. DOD Manual for Building a Technical Thesaurus, Office of Naval Research, Project FX, April 1966. (AD-633279).

48. Herner, Saul. The Relationship of Information Use Studies and the Design of Information Storage and Retrieval Systems, Herner and Company, Washington, D. C., December 1958.

49. Herner, Saul. System Design Evaluation and Costing--in Plain English, Herner and Company, Washington, D. C., March 1967. (AD 657-788).

50. Herner, Saul; Herner, Mary. "Information Needs and Uses," In: Annual Review of Information Science and Technology, Volume 2, John Wiley and Sons, New York, 1967.

51. Holzbaur, F. W.; Farris, E. H. Library Information Processing Using an On-Line, Real-Time Computer System, International Business Machines Corp., Systems Development Div., Poughkeepsie, New York, December 7, 1966.

52. Hooper, Robert S. Development and Application of a Technique for the Evaluation of Document Retrieval Systems, American University, Washington, D. C., May 1967.

53. Ide, Eleanor; Williamson, R.; Williamson, D. "The Cornell Programs for Cluster Searching and Relevance Feedback." Information Storage and Retrieval, June 1967.

54. Informatics, Inc. Display Oriented Computer Usage System, Interim Technical Report, October 1964 - April 1966. Bethesda, Maryland, June 1966.

55. Jones, Paul E.; Curtice, Robert N. "A Framework for Comparing Term Association Measures." American Documentation, 18: 153-161, July 1967.

56. Jones, Paul E.; Giuliano, V. E.; Curtice, R. M. Papers on Automatic Language Processing, 3 Volumes, Arthur D. Little, Inc., Cambridge, Massachusetts, February 1967.

57. Jones, Sparck K. Semantic Markers, Cambridge Language Research Unit, Cambridge, England, November 1965.

58. Jones, Sparck; Jackson, David. "Current Approaches to Classification and Clump Finding at the Cambridge Language Research Unit." Computer Journal, 10: 29-37, May 1967.

59. Jones, Sparck; Jackson, David. The Use of the Theory of Clumps for Information Retrieval, Cambridge Language Research Unit, Cambridge, England, June 1967.

60. Katter, Robert V. Experimental Investigations of a Method for Analyzing Document Representation: Progress Report, System Development Corp., Santa Monica, California, August 9, 1966.

61. Kellogg, Charles H. Converse - A System for the On-Line Description and Retrieval of Structured Data Using Natural Language, System Development Corp., Santa Monica, California, May 26, 1967.

62. Kellogg, Charles H. On-Line Translation of Natural Language Questions into Artificial Language Queries, System Development Corp., Santa Monica, California, April 28, 1967.

63. Kent, Allen. Textbook on Mechanized Information Retrieval, 2nd Edition, Inter-Science Publishers, New York, 1966.

64. Kessler, M. M. "The MIT Technical Information Project." Physics Today, 18: 28-36, March 1965.

65. Kilgour, Frederick G. "Systems Concepts and Libraries." College and Research Libraries, 28: 167-170, May 1967.

66. King, Donald W. "Design and Evaluation of Information Systems," In: Annual Review of Information Science and Technology, Volume 3, Encyclopedia Britannica, Inc., Chicago, 1968.

67. King, D. W. Evaluation During File Development of the Glass Technology Coordinate Index File, for the U. S. Department of Commerce Patent Office, Westat Research, Inc., Bethesda, Maryland, November 1967.

68. King, D. W.; Isakov, P. "Preliminary Evaluation of the Glass Technology Coordinate Index File." Presented at the 7th Annual Meeting of the Committee for International Cooperation in Information Retrieval among Examining Patent Offices, Stockholm, Sweden, September 1967.

69. Kline, Melvin B.; Lifson, Melvin W. "Systems Engineering," In: Cost Effectiveness, John Wiley and Sons, New York, 1968.

70. Kraft, Donald H. "A Total Systems Approach to Library Mechanization." Texas Conference on Library Mechanization, 1: 23-24, March 1966.

71. Lancaster, F. W. "Evaluating the Performance of a Large Operating Retrieval System," In: Electronic Handling of Information: Testing and Evaluation, Academic Press, London, 1967.

72. Lancaster, F. Wilfrid. Information Retrieval Systems, John Wiley and Sons, New York, 1968.

73. Leimkuhler, Ferninand F. "A Literature Search Model." Bulletin of the Operations Research Society of America, 15, Spring 1967.

74. Lesk, M. E.; Salton, Gerald. "Design Criteria for Automatic Information Systems." Information Storage and Retrieval, June 1966.

75. Levien, R. E.; Maron, M. E. "A Computer System for Inference Execution and Data Retrieval." Communications of the ACM, 10: 715-721, November 1967.

76. Levien, Roger; Maron, M. E. Relational Data File: A Tool for Mechanized Inference Execution and Data Retrieval, Rand Corporation, Santa Monica, California, December 1965.

77. Licklider, J. C. R. Libraries of the Future, MIT Press, Massachusetts Institute of Technology, Cambridge, Massachusetts, 1965.

78. Lipentz, Ben-Ami. "Information Storage and Retrieval." Scientific American, 215: 224-232, September 1966.

79. Liston, D., Jr. "Information Systems: What They Do, How They Work." Machine Design, 38: 190, July 21, 1966.

80. Lowe, Thomas C. Design Principles for an On-Line Information Retrieval System, University of Pennsylvania, Moore School of Electrical Engineering, Philadelphia, December 1966.

81. Marron, H.; Snyderman, M. "Cost Distribution and Analysis in Computer Storage and Retrieval - 2." American Documentation, 18: 162-164, July 1967.

82. Mayer, Steven E. A Model for the Description of Information Retrieval Systems, Preliminary Report 2, Human Sciences Research, Inc., McLean, Virginia, December 1965.

83. McCarn, Davis B.; Stein, C. R. "Intelligence Systems Evaluation," In: Electronic Handling of Information: Testing and Evaluation, Academic Press, London, 1967.

84. McDonough, A. M. Information Economics and Management Systems, McGraw-Hill, New York, 1963.

85. Meadow, Charles T. The Analysis of Information Systems: A Programmer's Introduction to Information Retrieval, John Wiley and Sons, New York, 1967.

86. Melton, Jessica S. "Automatic Language Processing for Information Retrieval: Some Questions," In: Proceedings of the 1966 ADI Annual Meeting, Adrianne Press, Woodland Hills, California, 1966.

87. Minsky, Marvin L. "Artificial Intelligence." Scientific American, 215: 246-260, September 1966.

88. Moore, Edythe. "Systems Analysis: An Interview." Special Libraries, 58: 87-90, February 1967.

89. Murdock, John W.; Liston, D. M. "A General Model of Information Transfer: Theme Paper 1968 Annual (ADI) Convention." American Documentation, 18: 197-208, October 1967.

90. National Science Foundation, Office of Science Information Service. Current Research and Development in Scientific Documentation, No. 14, Washington, D. C., 1966.

91. Nelson, Edward A. Management Handbook for the Estimation of Computer Programming Costs, System Development Corp., Santa Monica, California, March 1967.

92. Oliver, Lawrence H.; Mitchell, Claude; Fitzpatrick, Edmund W.; Jacobson, Herbert. An Investigation of the Basic Processes Involved in the Manual Indexing of Scientific Documents, General Electric Co., Information Systems Operation, Rethesda, Md., February 11, 1966.

93. Optner, Stanford L. Systems Analysis for Business and Industrial Problem Solving, Prentice-Hall, Englewood Cliffs, New Jersey, 1966.

94. Perry, James W. Relevance in Document Selection, Some Fundamental Considerations, Systems Engineering Department, University of Arizona, Tucson, August 20, 1965.

95. Perry, James W.; Kent, Allen. Tools for Machine Literature Searching, Interscience Publishers, Inc., New York, 1958.

96. Perry, James W.; Kent, Allen; Berry, Madeline M. Machine Literature Searching Interscience Publishers, Inc., London, 1956.

97. Pollock, Stephen. "Evaluation and Comparison of Information Retrieval Systems." Presented at the 31st National Meeting of the Operations Research Society of America, New York, June 1, 1967.

98. Prince, Thomas R. Information Systems for Management Planning and Control Richard D. Irwin, Inc., Homewood, Ill., 1966.

99. Quade, E. S. Some Problems Associated with Systems Analysis Rand Corp, Santa Monica, California, June 1966.

100. Rees, Alan M.; Schultz, D. G. A Field Experimental Approach to the Study of Relevance Assessments in Relation to Document Searching. Center for Documentation and Communication Research, Case Western Reserve University, Cleveland, Ohio, 1967.

101. Ross, Douglas T. "The AED Free Storage Package." Communications of the A.C.M. 10: 481-492, August 1967.

102. Salton, Gerald. "A Document Retrieval System for Man-Machine Interaction." In: Association for Computing Machinery, Proceedings of the 19th Annual Conference, Philadelphia, 1964.

103. Salton, Gerald. "The SMART Project - Status Report and Plans." Information Storage and Retrieval, June 1967.

104. Salton, Gerald; Lesk, M. E. "Information Analysis and Dictionary Construction." Information Storage and Retrieval, June 1966.

105. Saracevic, T.; Rees, A. Towards the Identification and Control of Variables in Information Retrieval Experimentation. Comparative Systems Laboratory Technical Report No. 6, Center for Documentation and Communication Research, Western Reserve University, January 1966.

106. Schecter, George. Information Retrieval - A Critical Review. Academic Press, London, 1967.

107. Sharp, John R. "Content Analysis, Specification, and Control." In: Annual Review of Information Science and Technology, Volume 2, John Wiley and Sons, New York, 1967.

108. Sharp, John R. "The SLIC Index." American Documentation, 17: 41-44, January 1966.

109. Shaw, R. R. "Information Retrieval." Science, 140: 606-609, May 10, 1963.

110. Smith, J. L. Micro: A Strategy for Retrieving, Ranking, and Qualifying Document References. System Development Corp., Santa Monica, California, January 15, 1966.

111. Spiegel, J.; Summers, J. K.; Bennett, A. M. Aesop: A General Purpose Approach to Real-Time, Direct Access Management Information Systems. Mitre Corp., Bedford, Massachusetts, June 1966.

112. Stevenson, C. G. "Checklist for Review and Evaluation of Technical Libraries." Special Libraries, 58: 106-110, February 1967.

113. Swanson, Rowena. Information Sciences, 1965. Annual Report No. 3, 1965, Air Force Office of Scientific Research, Directoriate of Information Sciences, Washington, D. C., January 1966.

114. Swanson, Rowena. Information System Network...Let's Profit from What We Know. Office of Aerospace Research, U. S. Air Force, June 1966.

115. Swets, John A. Effectiveness of Information Retrieval Methods. Bolt Beranek and Newman, Inc., Cambridge, Massachusetts, June 15, 1967.

116. Tague, Jean. "An Evaluation of Statistical Association Measures." In: Proceedings of the 1966 ADI Annual Meeting, Adrianne Press, Woodland Hills, California, 1966.

117. Tanning, Edward A. Operations of a Document Retrieval System Using a Controlled Vocabulary. Dayton University Research Institute, Ohio, March 1966.

118. Taulbee, Orrin E. "An Approach to Comprehensive Evaluation." In: Electronic Handling of Information: Testing and Evaluation, Academic Press, London, 1967.

119. Taulbee, Orrin E. "Content Analysis, Specification, and Control." In: Annual Review of Information Science and Technology, Volume 3, Encyclopedia Britannica, Inc., Chaciago, 1968.

120. Taulbee, Orrin E. "Model of Document Retrieval." Library Science, 3: 248-258, 1966.

121. Tukey, John W. "A Statistican's Comment." In: Electronic Handling of Information: Testing and Evaluation, Academic Press, London, 1967.

122. Ullmann, Hans C. Indexing Techniques. System Development Corp., Santa Monica, California, February 23, 1966.

123. Weinwurm, George F. On the Economic Analysis of Computer Programming. System Development Corp., Santa Monica, California, August 1967.

124. Wiederkehr, R. R. V. Relationships Between the Common Cost-Benefit Measures. Progress Report Submitted to National Science Foundation, November 1967.

125. Wilcox, Richard H. Information Systems Summaries, 3rd Edition, Office of Naval Research, Dept. of the Navy, Washington, D. C., July 1965.

126. Williams, J. H. Discriminant Analysis for Content Classification. IBM Federal Systems Division, Bethesda, Maryland, February 1966.

127. Wilson, Ira G.; Wilson, Marthann E. Information, Computers and System Design. John Wiley and Sons, New York, 1965.

128. Wyllys, Ronald E. How Should a Document-Handling System Be Designed? University Libraries and Computer Sciences Dept., University of Wisconsin, Madison, 1966.

129. Zunde, Pranas; Armstrong, F. T.; Stretch, T. T. "Evaluating and Improving Internal Indexes." In: Proceedings of the 30th Annual Meeting of the American Documentation Institute, New York, October 1967.